Researching Less-Resourced Languages – the DigiSami Corpus

Kristiina Jokinen
University of Helsinki, Finland and AIRC, AIST Tokyo Waterfront, Japan
kristiina.jokinen@helsinki.fi

Abstract

Increased use of digital devices and data repositories has enabled a digital revolution in data collection and language research, and has also led to important activities supporting speech and language technology research for less-resourced languages. This paper describes the DigiSami project and its research results, focussing on spoken corpus collection and speech technology for the Fenno-Ugric language North Sami. The paper also discusses multifaceted questions on ethics and privacy related to data collection for less-resourced languages and indigenous communities.

Keywords: corpus collection, under-resourced languages, North Sami

1. Introduction

Several projects and events have increased research activities for under-resourced languages during the past years. For instance, the DLDP-project (Digital Language Diversity Project) is to advance the sustainability of Europe's regional and minority languages, while the Flare-net network and the LRE Map (Calzolari et al. 2012) have had a big impact on sharing language resources and making speech corpora freely available. However, there are many challenges that researchers and developers face when aiming at the same technology standards for less-resourced and endangered languages as those for majority languages. The challenges do not only concern scarce data and non-optimal algorithms, but also issues inherently related to the cultural contexts of linguistic communities in general, and shared background of minority cultures in particular. It is thus important to pay attention to community-based techniques in data collection and technology development (crowdsourcing in a wide sense) as well as trying to connect the communities with other small language communities (Soria et al. 2013). An important aspect of such work is to empower minority language speakers with the knowledge and skills to create and share content for digital devices using their own language.

The DigiSami project (http://www.helsinki.fi/digisami/) is supported by the Academy of Finland in a wider context of a joint initiative with the Hungarian Academy of Sciences. The focus of the Digital Citizens framework project was to support collaborative research on endangered Fenno-Ugric Languages and to develop tools and resources for automatic language processing as well as to experiment with new technology applications. The main motivation was to improve digital visibility and viability of the target languages, and to explore different choices for encouraging and maintaining the use of less-resourced languages in the digitalized world. The goals of the DigiSami project are discussed in Jokinen (2014) and Jokinen et al. (2017).

Figure 1. The Sami language areas at the beginning of the 20th century. Abbreviations: So - South Sami, Um - Ume Sami, Pi - Pite Sami, Lu - Lule Sami, No - North Sami, In - Inari Sami, Sk - Skolt Sami, Ak - Akkala Sami (extinct), Ki - Kildin Sami, Tr - Ter Sami. From Jokinen et al. (2017), original source Sammallahti (1998).

The DigiSami project deals with the North Sami language (Davvisámegiela) which belongs to the Fenno-Ugric language family and is one of the nine Sami languages spoken in the northern part of Europe: Scandinavia, Finland and the Kola Peninsula in Russia (Seurujärvi-Kari et al. 1997). Figure 1 shows the Sami languages and their geographical distribution. There are about 40 000 speakers of the Sami languages, and of these about half speak North Sami, which has gained the status of a lingua franca. North Sami also enjoys official status in Norway and Finland, and has language technology tools so it is possible to develop online content analysis and interactive applications (see more in Jokinen et al. 2017).

2. The DigiSami Corpus

The project organised data collection in three locations in Finland and two locations in Norway, see Figure 2. The locations were selected to include representative locations of the variations of North Sami spoken in the Sápmi area: the border between the countries of Norway and Finland divides the area in two, and there is also a major language division between the Eastern and Western dialects of North Sami, so that the Sami language spoken in Kautokeino and Enontekiö belong to Western North Sami, while the others belong to Eastern North Sami. In Inari, Inari Sami is spoken as a separate language (Sammallahti 1998). See Jokinen (2014) and Jokinen and Wilcock (2014) for more about data collection.

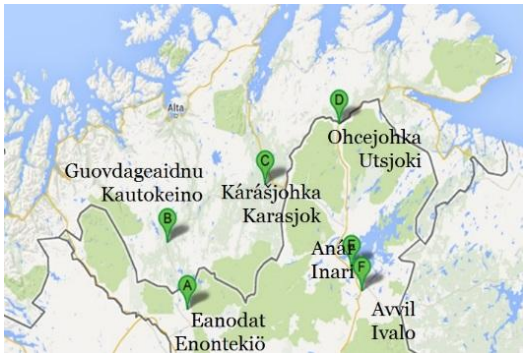

Figure 2. The data collection locations in Norway (Kautokeino, Karasjok) and in Finland (Ivalo, Inari, Utsjoki). No data was collected in Enontekiö although originally intended.

The DigiSami corpus contains two types of spoken data: video data on conversational spontaneous speech, and read speech from participants reading out Wikipedia articles. Participants were bilingual (North Sami and either Finnish or Norwegian), and had lived most of their life in the Sápmi area. Although local dialects in the participants' speech cannot be tracked due to the small number of participants, the corpus contains notable pronunciation variation between Norwegian and Finnish participants. This has been further studied by Jokinen et al. (2016a) using automatic recognition methods for speaker and dialect identification.

The dialogue corpus has been transcribed using Praat, and annotated following the MUMIN guidelines set out in Allwood et al. (2007) and Paggio et al. (2010). Multimodal annotation concerns especially laughing, body movements and topics for multimodal conversation studies.

Basic statistics of the corpus are given in Table 1, and demographic facts in Figure 3.

	Read speech		Conversational speech	
Dialects	#Spkr	Duration (hrs)	#Spkr	Duration (hrs)
Kautokeino	4	1.03	-	-
Karasjoki	7	0.72	6(1)	1.5
Ivalo	7	0.72	7(1)	0.72
Utsjoki	6	1.07	6(1)	1.03
Inari	4	0.73	-	-
Total	**28**	**3.36**	**19**	**4.28**

Table 1: Basic Statistics of the DigiSami Corpus

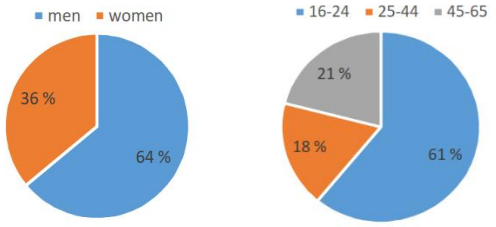

Figure 5 The participant's gender (left) and age (right).

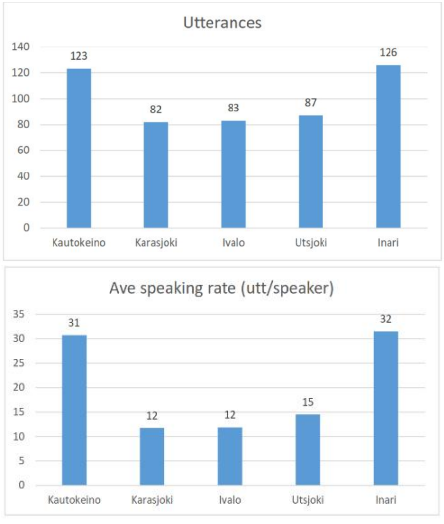

Figure 6. The number of utterances and the average speaking rate in the DigiSami corpus for each of the five locations. The Kautokeino and Inari numbers refer to read speech, while the Karasjok, Ivalo, and Utsjoki show conversational data.

The conversational corpora are comparable in that the speakers seem to talk at a similar rate and produce a similar number of utterances. Conversations take place among young adults at schools, and apparently reflect similarities among the young Sami people in Norway (Karasjok) and Finland (Ivalo and Utsjoki).

From the questionnaire that the participants filled in before the recordings, we can also find out that most participants used North Sami as the main communication means when interacting with other people, a third of the participants spoke North Sami when communicating with relatives on either mother's or father's side only, and only three (11%) of the participants used North Sami at official places and at work, but not at home (Figure 3). Communication context for speaking North Sami was predominantly at home, but also at work (or school), and to a lesser degree at public places like shops, offices, and restaurants (Figure 4). The percentages in this figure do not add up to 100%, since the participants could mark as many alternative locations as they wanted, and the categories are not mutually exclusive either. More descriptive data analysis can be found in Jokinen (2014).

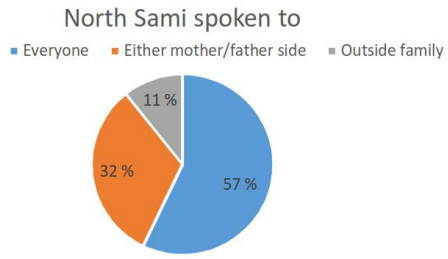

Figure 3 Reported interlocutors when speaking North Sami.

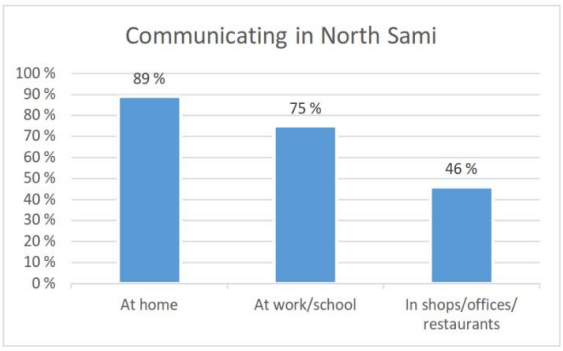

Figure 4 Communication contexts when speaking North Sami among the participants.

The corpus has been widely used in various studies on spoken interaction. The pronunciation differences are investigated in Jokinen et al. (2016a), while Grönroos et al. (2016) present morphological segmentation for the North Sami language using the active learning method. Trong et al. (2018) use an end-to-end dialect recognition system based on the deep learning method and discuss its use as enabling technology for building interactive applications. Hiovain and Jokinen (2016) discuss laughter types and Trong et al. (forthcoming) study relation between laughter, topics and body movement, and also compare the corpus with related corpora in Finnish and Estonian.

The papers are accessible through the project website, while the DigiSami corpus is available via the CSC website by contacting the author.

3. SamiTalk

The DigiSami project also aims at technology applications that would make the North Sami more visible in the digital world and also boost the North Sami status as a prestige language, usable for new digital applications. The project thus designed SamiTalk, a human-robot interactive application for accessing North Sami Wikipedia (Wilcock et al. 2017). SamiTalk is based on the WikiTalk and MoroTalk applications (Jokinen and Wilcock 2013, Wilcock 2012) which allow the users to search for interesting information from Wikipedia and interactively chat with the system. Although more work is needed to develop the Sami speech technology components further, it is expected that in the near future SamiTalk type applications can be effectively used for tasks such as interactive language learning and preserving cultural heritage via storytelling, besides the common information retrieval or question-answering tasks. However, these developments will also need effort and active participation from the Sami people to proceed in a co-design manner towards applications that also respect the fragile cultural context (see Section 4).

During the data collection events, also a small number of North Sami Wikipedia articles was produced to encourage the community to develop North Sami Wikipedia further. Wikipedia statistics (http://stats.wikimedia.org/) tells that the North Sami Wikipedia started in 2004, and ranks in the middle of all Wikipedias with over 7500 articles. Currently only North Sami has a Wikipedia, but during the data collection sessions, interest was also sparkled within Inari Sami community to start a Wikipedia of their own. The number of North Sami page requests is about half a million page requests per month, but it is not possible to know how many human readers there are. In general, however, an accelerating circle can be noted: the more readers, the more editors who create more content, and the more content, the more readers. As for the content, North Sami Wikipedia is 134th of about 300 in number of articles. Articles are fairly short, about 500 characters long on average, but as wikis grow, the average article length grows as well. However, it there are many articles about towns around the world, since such articles can be easily created following a regular pattern with the town specific numbers filled from a database. Some editors automate the creation of articles but no statistics exists about how many articles have been created by translating them from some other Wikipedia.

The aim in DigiSami was not only to create more articles for SamiTalk, but to encourage development of a new public space that indigenous people could use to make their own voice heard through the information that the people themselves have created for a wider audience. Wikipedia has a well-suited format for documenting indigenous people's life and important events, places and people, since it is a form of public and non-commercial information technology. It is collectively produced and jointly developed and can thus also help to make the language and culture more visible by creating interest in the topics that the indigenous writers choose. Since much of the language technology research and common information retrieval technology uses Wikipedia as a resource, extending the existing North Sami Wikipedia is not only useful for the North Sami speakers, but strengthens the North Sami presence in digital world.

4. Ethical Aspects

The DigiSami corpus follows the standard rules, principles, and guidelines for ethical and privacy issues in data collection provided by Finland's National Advisory Board on Research Ethics. Each participant, or the parent of an underage participant, signed a consent form to take part in the corpus collection, and the research papers on data collection were sent to the local coordinators. The names, ages, and other demographic information that would give away personal identification of the participants are not present in the transcriptions, and research examples are chosen so that they do not identify individual speakers (see discussion of the guidelines in Jokinen 2011). Local participants were also invited to the SamiTalk demo at the IWSDS 2016 conference held in Saariselkä in Lapland (http://www.wikicfp.com/cfp/servlet/event.showcfp?eventid=47125&copyownerid=66338), and the book edited on the basis of the presentations also acknowledges the DigiSami project and its goals (Jokinen and Wilcock 2017).

Moreover, the DigiSami project has paid attention to the questions concerning data gathering from less-resourced languages in general. Confidentiality, privacy and respect are relevant issues in data collection, and issues on ethics and privacy are particularly important when collecting data from indigenous communities. These are multifaceted questions related to sensitive issues of the rights and ownership of the indigenous people, and a major concern is the people's rights and ownership of their own cultural heritage which should be respected. Since the culture and traditions are carried by the language, there is always an inherent element of cultural knowledge in the topics and the use of the language, and although corpus collection can aim at a neutral goal of documenting a less-resourced language, in practice it may not be possible to collect a culturally "neutral" corpus. Thus attention should be paid to the culture and tradition. The following questions are relevant in data collection in general, with special emphasis in the context of less-resourced and indigenous languages:

- What kind of data is gathered and what kind of activities are included? Written and spoken discourse already available publicly in books, TV radio, and films, is of different nature than recordings of private, spontaneous, and everyday behaviours, or open data available in internet and social media channels. However, as discussed by Oskal (2008), social and cultural research is not an individualistic process, but includes interaction with community members who produce knowledge in collaboration with the other members.
- Who owns the data? The rights to use the data can vary according to the copyright legislation, but there are also sensitive and complex issues related to Civil and Political Rights. An overview of the issues in relation to Sami languages can be found in Kokko (2010).
- What purpose is the data used for? The collection of data may be used for studying characteristics of an individual, a group, or a culture, in order to preserve their characteristics or to better understand their activities. The question also concerns how to define the appropriate ways to use the data, who can use it and who has the right to access the data.
- What kind of personal information is included? Text, speech, and video include different degrees of personal information and thus identification of the individuals

who produced the data can vary (Jokinen 2011). However, given present-day techniques of data processing, individual characteristics may be retrieved easily. On the other hand, distinction between the categories of personal vs collective may not appear so clear-cut: if the individuals' experience of their own activity is seen as an inherent part of the group activity, it equates with a collective view of the community's life and culture. The question of privacy thus does not only concern a consent from an individual but from a whole community, as the activity transfers from a single participant to the whole community of which the individual is a member. An important question is what kind of impact the corpus has on the participants' life.

Concerning the Sami culture and Sami languages, these issues have been actively investigated by existing cultural and social institutions in the area. For instance, Sámi Parliament, Sámi Museum Siida, Giellagas Institute and the Saami Culture Archive at the University of Oulu, and the Law School of University of Lapland, are among the main actors studying legal issues and ethical questions concerning the Sami culture and heritage.

5. Conclusions

The DigiSami corpus is available and has already been used in various research activities (see Section 3). Future steps could include data and knowledge collection by the people themselves and documenting this in Wikipedia articles or in other common digital archive formats. This could be organised in community halls as part of community activities, or in schools as part of practising Sami language writing in mother tongue classes. An important part of the activities is that they should arise from the language community and from the speakers' willingness and concern to work on the language themselves, rather than being seen as an outside activity that meets resistance. As mentioned above, applications such as SamiTalk have potential to be used in language teaching as well as assisting story-telling for culture preservation. Examples of using technology to preserve the culture already exists (e.g. Rodil 2014), and a great example is the way game apps have been used successfully in Australian aboriginal languages (Aboriginal Australian language video game).

More information is also needed about the specific needs for digital information and applications that the indigenous users may find useful. In spring 2016 the DLDP consortium conducted a survey and gathered information about the personal digital use of the language and about any known digital resource and services that make use of the language.

Improving viability in the digital world as well as revitalization of the language use is more likely to succeed if there are local people actively involved in the process. Collection of corpora for further language technology studies requires that certain standards for the development for language technology tools and applications are taken into account. The DigiSami project has taken steps in this direction. It is hoped that the DigiSami corpus will prove useful as the first systematic collection of multimodal North Sami conversation for research in speech and interaction technology, as well as for cultural studies, and that it can also support revitalisation and digital viability of the Sami languages and Sami culture in general.

6. Acknowledgements

The work has been carried out in the Academy of Finland project Fenno-Ugric Digital Citizens (grant n°270082). The author would like to thank all the project workers, students, participants, teachers, and colleagues who have been involved in the project, assisted and contributed to the work in various stages of the project.

7. Bibliographical References

Allwood, J., Cerrato, L., Jokinen, K., Navarretta, C. and Paggio, P. (2007). The MUMIN Coding Scheme for the Annotation of Feedback, Turn Management and Sequencing. In Martin et al. (eds) *Multimodal Corpora for Modelling Human Multimodal Behaviour*.

Boersma, P. and D. Weenink (2009). Praat: doing phonetics by computer (version 5.1.05). Retrieved May 1, 2009, from http://www.praat.org/

Calzolari N., Del Gratta R., Fracopoulo G., Mariani J., Rubino F., Russo I. and Soria C. The LRE Map. Harmonising Community Description of Resources. *Proceedings of LREC 2012*, Istanbul, Turkey, pp. 1084—1089, 2012.

Grönroos, S.A., Hiovain, K., Smit, P., Rauhala, I., Jokinen, K., Kurimo, M., and Virpioja, S. (2016). Low-Resource Active Learning of Morphological Segmentation. Northern European Journal of Language Technology, Vol. 4, pp 47–72, DOI 10.3384/nejlt.2000-1533.1644

Grünthal, R., Siegl, F.: Uralilaisten kielten pensasmalli ja arvioidut puhujamäärät (The "bush model" of the Uralic languages and the estimated numbers of speakers) (2012). Department of Finno-Ugric Studies, University of Helsinki.

Hiovain, K. and Jokinen, K. (2016). Acoustic Features of Different Types of Laughter in North Sami Conversational Speech. LREC Workshop Just talking – casual talk among humans and machines, Portorož.

Jokinen, K. (2011) Multimodal Information - Collection and Analysis of Interactive Data. HCII 2011. Orlando, U.S.

Jokinen, K. (2014). Open-domain interaction and online content in the Sami language. In: Proceedings of Ninth International Conference on Language Resources and Evaluation (LREC 2014). Reykjavik.

Jokinen, K., Hiovain, K., Laxström, N., Rauhala, I., Wilcock, G. (2017). DigiSami and Digital Natives: Interaction Technology for the North Sami Language. In: Jokinen, K. and Wilcock, G. (Eds.). *Dialogues with Social Robots – Enablements, Analyses, and Evaluation.* pp. 3-19. Springer. http://www.springer.com/gb/book/9789811025846 DOI: 10.1007/978-981-10-2585-3.

Jokinen, K., Trong, T.N., Hautamäki, V. (2016). Variation in Spoken North Sami Language, Interspeech.

Jokinen, K. and Wilcock, G. (2017). *Dialogues with Social Robots – Enablements, Analyses, and Evaluation*. Lecture Notes in Electrical Engineering, Vol 427, Springer, DOI: 0.1007/978-981-10-2585-3

Jokinen, K. and Wilcock, G. (2014). Community-Based Resource Building and Data Collection. The 4th International Workshop on Spoken Language Technologies for Under-resourced Languages (SLTU'14). St.Petersburg, Russia.

Jokinen, K. and Wilcock, G. (2013). Multimodal Open-domain Conversations with the Nao Robot. In: Mariani,

J., Devillers, L., Garnier-Rizet, M. and Rosset, S. (eds.) *Natural Interaction with Robots, Knowbots and Smartphones - Putting Spoken Dialog Systems into Practice*. Springer Science+Business Media

Kokko, K.T. (Ed) (2010). Kysymyksiä saamelaisten oikeusasemasta. (”Issues Concerning the Legal Status of the Sámi people”) Lapin yliopiston oikeustieteellisiä julkaisuja Sarja B no 30 (University of Lapland Publications Series B Number 30), Rovaniemi. http://urn.fi/URN:NBN:fi:ula-201201161003

Oskal, N. (2008). The question of methodology in indigenous research. A philosophical exposition. In H. Minde, S. Jentoft, H. Gaski & G. Midré (eds.) *Indigenous peoples: self-determination, knowledge, indigeneity*. Eburon: Delft, pp. 331–345.

Paggio, P., J. Allwood, E. Ahlsén, K. Jokinen, C. Navarretta (2010). The NOMCO multimodal Nordic resource - goals and characteristics. In *Proceedings of LREC 2010*, 2968-2973.

Prószéky, G. (2011) Endangered Uralic Languages and Language Technologies. Proceedings of Language Technologies for Digital Humanities and Cultural Heritage Workshop, pages 1–2, Hissar, Bulgaria, 16 September 2011.

Rodil, K. (2014). A Participatory Perspective on Cross-Cultural Design. In Ebert, A., van der Veer G.C., Domik, G., Gershon, N.D., Scheler, I. (eds.). *Building Bridges: HCI, Visualization, and Non-formal Modeling*, Springer Lecture Notes in Computer Science, pp. 30-46.

Sammallahti, P. (1998). *The Saami Languages: An Introduction*. Davvi Girji, Kárášjohka.

Seurujärvi-Kari, I., Pedersen, S. and Hirvonen, V. (1997). The Sámi. The indigenous people of northernmost Europe. *European Languages 5*. Brussels: European Bureau for Lesser Used Languages.

Soria, C., Mariani, J. and Zoli, C. (2013). Dwarfs sitting on the giants' shoulders – how LTs for regional and minority languages can benefit from piggybacking major languages Proceeding of XVII FEL Conference 10/2013, Ottawa.

Trong, T.N., Jokinen, K., Hautamäki, V. (2018). Enabling Spoken Dialogue systems for low-resource languages – End-to-end dialect recognition for North Sami. Proceedings of the IWSDS 2018, Singapore.

Wilcock, G. (2012). WikiTalk: a Spoken Wikipedia-based Open-domain Knowledge Access System. Proceedings of the COLING-2012 Workshop on Question Answering for Complex Domains, Mumbai, pp. 57-69.

Wilcock, G., Laxström, N., Leinonen, J., Smit, P., Kurimo, M., Jokinen, K. (2017). Towards SamiTalk: a Sami-speaking Robot linked to Sami Wikipedia. In: Jokinen, K. and Wilcock, G. (Eds.). *Dialogues with Social Robots – Enablements, Analyses, and Evaluation*. pp. 343-351. Springer. http://www.springer.com/gb/book/9789811025846 DOI: 10.1007/978-981-10-2585-3.

8. Language Resource References

Aboriginal Australian language video game. An interview of the developers. The DLDP Newsletter #1 - December 2016. http://mpegmedia.abc.net.au/radio/local_canberra/audio/201609/abn-2016-09-27-video-game-offers-new-life-to-ancient-indigenous-language.mp3 [Online; accessed 25-September-2017]

DigiSami Project (2017). Academy of Finland project http://www.helsinki.fi/digisami/

DigiSami Corpus (2017). Available from the project leader, and accessible through Center for Scientific Computing. www.csc.fi

Digital Language Diversity Project. Erasmus+ Programme Grant Agreement No. 2015-1-IT02-KA204-015090 http://www.dldp.eu/ [Online; accessed 25-September-2017]

ESFRI. European Strategy Forum on Research Infrastructures. http://ec.europa.eu/research/infrastructures/index_en.cfm?pg=esfri [Online; accessed 25-September-2017]

Giellagas Institute. (2015). The Saami Culture Archive of University of Oulu. http://www.oulu.fi/giellagasinstitute/the_saami_culture_archive. [Online; accessed 25-September-2017]

Simple4All Consortium. (2015). Simple4All: developing automatic speech synthesis technology http://simple4all.org/ [Online; accessed 25-September-2017]